DOCUMENT RESUME

ED 270 465 TM 860 343

AUTHOR Capie, William; Cronin, Linda
TITLE How Many Teacher Performance Criteria Should There Be?
PUB DATE Apr 86
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Correlation; Decision Making; *Evaluation Criteria; Field Tests; Generalizability Theory; Junior High Schools; Merit Pay; *Minimum Competency Testing; Pretests Posttests; Regression (Statistics); Science Education; Scores; Teacher Certification; Teacher Effectiveness; *Teacher Evaluation; Teacher Qualifications; *Test Reliability; *Test Validity
IDENTIFIERS *Georgia; Group Assessment of Logical Thinking; Middle Grades Integrated Process Skill Test; Teacher Effectiveness Index; *Teacher Performance Assessment Instruments

ABSTRACT

This paper assesses the credibility of a single total instrument score and various logical sub-scores derived from a series of summative judgments about the quality of teaching performance. The objectives were to compare the generalizability of alternative Teacher Performance Assessment Instrument (TPAI) scores, to compare the dependability of decisions which could be made with the scores, and to compare the relationship of the scores with learner achievement. Measures were made of teacher performance using the revised version of the TPAI. Learner ability was assessed with the Group Assessment of Logical Thinking in order to equate classes. Learner achievement was assessed with the Middle Grades Integrated Process Skill Test. Results show that: (1) the aggregation of summative judgments used in the TPAI scoring can be a valid and reliable procedure; (2) intermediate levels of scoring such as the TPAI competencies are more desirable than total instrument scores; (3) the total is a more reliable, but less valid, indicator of effectiveness; and (4) validity and dependability coefficients are adequate evidence to support the validity and reliability of the competency scores. Caution should be excerised in inferring causality of these teacher behaviors or learner outcomes based on these results. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED270465

HOW MANY TEACHER PERFORMANCE CRITERIA SHOULD THERE BE?

William Capie
Linda Cronin

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. Capie

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Teacher Assessment Project
The University of Georgia
Athens, GA 30602

A paper presented at the annual meeting
of the
American Educational Research Association
San Francisco, CA
April 18, 1986

TM 860 343

HOW MANY TEACHER PERFORMANCE CRITERIA SHOULD THERE BE?

A query such as the one posed in the title has a simplistic ring to it. It implies that there is an answer such as 2,143 or 2. More importantly, it begs the larger question "criteria for what?" In Georgia, the answer has been "criteria for initial certification" since the fall of 1980 when performance based certification was initiated. Of course, other states have many purposes for the assessment systems that have been designed and/or implemented since 1980. In addition to Georgia, states which have pursued assessment for initial certification include Virginia, Mississippi and Florida. Florida, Tennessee and Texas have been notable among states which have attempted to differentiate teachers for merit pay or career ladders based on on-the-job assessment among other criteria.

Assessment for merit pay may imply a different kind of decision-making process than does assessment for initial certification. While certification suggests that there will be a determination that the candidate possesses a set of basic skills, each of which will be assessed, merit pay might be based on a single decision about the quality of the teaching performance(s) assessed. However, even merit decisions could be based on a number of skill areas, lest a satisfactory over-all score be earned with a strong performance in one area compensating for a weak performance in another. Initial policy in the Texas career ladder program reflected this kind of concern with the stipulation that "Instructional Strategies" be weighted twice that of other areas such as "Classroom Management."

The nature of the assessment instrument is a strong determinant in resolving the question "How many teacher performance criteria should there be?" Data collection systems such as the Florida Performance Measurement System or the assessment component of the Beginning Teacher Assistance Program in Virginia are quite complex and, to date, not well-documented. A system of summative judgments, such as the Teacher Performance Assessment Instruments or the Teacher Assessment and Development System, produces a set of scores which can be aggregated in a number of ways which can reflect different needs in an assessment context. Capie (1983) has described a number of these alternative data aggregation possibilities and has speculated about their effects on assessment results.

The bottom line in designing an assessment system is that the scores which are generated must be credible, whether they are derived from an event sampling procedure such as the F.P.M.S., from a time sampling procedure such as that used in Virginia, or from an aggregated checklist of summative judgments such as the T.P.A.I. Credibility of teacher performance measures has a number of dimensions ranging from common sense believability to psychometric concerns such as validity and reliability. (Capie, 1985) He points out that concern for reliability must include both minimizing observer error and maximizing the ability to differentiate candidates from each other or from a criterion. Hambledon (1984) has described a number of aspects of the validity of criterion referenced test scores.

PURPOSE

The purpose of this paper is to assess the credibility of a single total instrument score and various "logical" sub-scores which are derived from a series of summative judgments about the quality of teaching performances. Specifically, the objectives were to compare the generalizability of alternative T.P.A.I. scores, to compare the dependability of decisions which could be made with the scores, and to compare the relationship of the scores with learner achievement.

CONTEXT

The Teacher Performance Assessment Instruments (TPAI) were developed between 1976 and 1980 and were introduced for initial certification in Georgia beginning in 1980. Beginning in 1981, a revision of the T.P.A.I. was undertaken. Multiple purposes were served by the revision including expansion of instrument content, improving feedback to teachers, and improving the psychometric qualities of the instrument. The revised T.P.A.I. was used for certification purposes in September 1985. The final field test of the revised T.P.A.I. provided an opportunity to conduct this and other research related to instrument design and use.

The site for the field test was selected in an effort to minimize problems which had threatened earlier T.P.A.I. validation efforts. Middle school science classes were selected since one concern during instrument development was the heavy reliance on teacher effectiveness research which was completed in limited settings (e.g. elementary mathematics and reading classes). Also, the science classes allowed for a common area of instruction

across all classes and, consequently, a common post-test which helped reduce random variation associated with these factors and helped ensure the curriculum relevance of the post-test. Finally, the topic for the unit was science process skills, a set of high cognitive level outcomes which are much different than the objectives of previous T.P.A.I. validation efforts which were restricted to teacher-made tests addressing relatively low-level cognitive outcomes.

Teachers were provided a two week unit plan during a ninety minute meeting when the nature of the unit and the requirements of the research were explained. The unit consisted of ten lessons, each with specified objectives and an outline of activities. The equipment that was required for the lessons was relatively common and easy to find. The activities were somewhat open in that teachers could modify them to suit the needs of their particular situation. Each of the lessons required some individual or group activity. The lessons were expected to be somewhat demanding for teachers, a fact which would increase variance in teachers' T.P.A.I. scores.

Instruments

Measures were made of teacher performance using the revised version of the T.P.A.I. Learner ability was assessed with the Group Assessment of Logical Thinking (GALT) (Roadrangka, Padilla and Yeany, 1983) in order to equate classes. Learner achievement was assessed with the Middle Grades Integrated Process Skill Test (MGIPT) (Cronin and Padilla, 1986).

The T.P.A.I. The revised version of the T.P.A.I. consists of eight teaching competencies each of which must be demonstrated

before a teacher is granted certification. Each competency is defined by three or four indicator statements, making a total of thirty indicators. Each indicator is defined by four descriptors, making a total of 120 descriptors. The competencies are the basic units for decision-making with the T.P.A.I. since each must be demonstrated satisfactorily prior to certification. The logic of the relationship among descriptors, indicators and competencies has been confirmed in an extensive content validation study. (Cronin and Capie, 1985) A listing of the eight competencies and their thirty constituent indicators is displayed in Figure 1.

INSERT FIGURE 1 ABOUT HERE

The more precise descriptor statements are the observation units which are used to derive the competency scores. A sample indicator with its defining descriptors is displayed in Figure 2. Observers examine all of the data which may be relevant to a descriptor and then make a decision about the adequacy of the observed performance in the context that is being assessed. The "relevant data" may be portfolio objectives for Indicator 1, tests and other assessment strategies for Indicators 5 and 6, or the entire observed lesson for Indicators 8-30. During a lesson, an observer will take notes to help recall important events from the lesson, but s/he will not be concerned with specific frequencies of certain behaviors. For example, the number of times that a teacher redirects off-task learners is not so important as whether or not the teacher attempts to redirect learners when necessary and whether or not these attempts were successful. Thus, the observer determines if there has been persistent off-task behavior

and, if so, determines how the teacher has dealt with it. These decisions, "acceptable" or "not acceptable," will be the scores for the descriptors in Indicator 28. The score for the indicator is derived from the descriptor scores using procedures that are specified in the instrument. Notice that the sample indicator has provision for giving the teacher credit for the indicator if there is no persistent off-task behavior. The indicator scores generated from the three observers' observation data are then aggregated to create the competency scores.

INSERT FIGURE 2 ABOUT HERE

The GALT. All students were administered the Group Assessment of Logical Thinking (GALT) (Roadrangka, Yeany and Padilla, 1983) as a pretest in an attempt to equate learner ability. GALT is a pencil and paper test consisting of 12 items that follows a double multiple choice format and requires approximately 30 minutes to administer. The test is designed to measure identification and control of variables., probabilistic, proportional, correlational, combinatorial, and conservational reasoning. There were two items for each mode of reasoning. The authors report a Cronbach's alpha estimate of .85.

The MGIPT. At the end of the unit the teachers all administered the same a post test (the Middle Grades Integrated Process Test, Cronbach alpha = .89) which was matched to the instruction, was composed of 40 multiple choice questions, and was provided by the researchers.

and the correlation between GALT and MGIPT (r=.62). For each learner the expected post-test score was subtracted from the observed post-test score. This difference was considered to be a teacher effect on the learner. The mean of these "teacher effects" for each class was used as a teacher effectiveness index. Thus, a variable reflecting class mean was available for use in subsequent analyses where classes were considered to be the sampling units.

Validity Indices. Simple correlations were computed between each of the TPAI measures and the teacher effectiveness index which was considered the criterion variable in the study. Thus, there were 122 correlations computed: 92 descriptors, 23 indicators, 6 competencies and a total instrument score. In addition a number of regression models were tested using selected indicator or competency scores as independent variables and the teacher effectiveness as a dependent variable.

Generalizability Analyses Generalizability theory was used to plan the analyses of the TPAI data. Four factors were identified as important sources of variation: teachers, individual observers, observer types and performance indicators. The four facet design, with individual observers nested within observer types, is arithmetically identical to a simpler three facet fully-crossed design with teachers, observer types and indicators as sources of variation. As a consequence, the simpler three facet model was used in the analyses of the competencies and the total instrument. Analyses of the indicators were completed with a two facet design. No generalizability analyses were completed for the individual descriptors. For each analysis, teachers were

considered to be the facet of differentiation and other facets were treated as random facets of generalization. Values of rho squared and phi(lambda) were computed to assess the suitability of the scores for differentiating teachers from each other and from the standard of having all indicators at or above the minimum level.

Results

Mean scores for the descriptors, indicators and competencies are included in Tables 1 and 2. Mean scores for descriptors could range from 0 to 1.00 to indicate the portion of observers who gave credit for each. Means for descriptors ranged from .23 (Descriptor 10c, Displays create a pleasant atmosphere) and .33 (Descriptor 29c, Learners are provided verbal feedback about acceptable behavior) to 1.00 (for a number of different descriptors).

Insert Tables 1 and 2 about here

Mean scores for indicators ranged from .38 (Indicator 24, Stimulates learner interest) to .95 (Indicator 13, Uses acceptable oral expression). Mean scores for the competencies ranged from .62 (Competency 6, Demonstrates appropriate instructional methods and Competency 7, Maintains a positive learning environment) to .90 (Competency III, Demonstrates acceptable written and oral expression and knowledge of the subject).

The mean of the teacher effectiveness index was -.08, indicating that, on average, the classes did about as well as their ability would have predicted. However, the variance was

large (standard deviation=3.2) with one class scoring nearly nine points (out of 36 possible) lower than predicted and another class performing nearly nine points higher than predicted based on their abilities.

Validity coefficients for descriptors are displayed in Table 3. They ranged from a low of -.26 (Descriptor 14a, Demonstrations and/or information presented to learners is accurate and up-to-date), to a high of .56 (Descriptor 10a, The classroom is free of litter). Twenty-three of ninety-two (one-fourth) were statistically significant ($p<.05$) and fourteen were greater than .30.

Insert Tables 2 and 3 about here

Validity coefficients for indicators and competencies are displayed in Table 4. For indicators the values ranged from .03 (Indicator 14, Demonstrates command of the school subject being taught) and .09 (Indicator 15, Gives explanations related to lesson content) to .53 (Indicator 10, Provides a physical environment that is conducive to learning). Eight of the twenty-three were statistically significant ($p<.05$) and four of these exceeded .30. Correlations between competency scores and the teacher effectiveness index were statistically significant for five of the six competencies, with the sole exception being Competency V (Communicates with Learners).

Insert Table 4 about here

The regression analyses of the indicators produced a number of meaningful models, one of which is summarized in Table 5. The four indicators which were used in the model accounted for 46.3 percent of the variance in between class differences in

achievement. The indicators dealt with the physical learning environment (#10), assessing learner progress (#11), using acceptable written expression (#12), and providing feedback to learners (#18). Similar analyses with the competency scores as independent variables were less able to predict class differences in achievement. The three variable model which is summarized in Table 6 accounted for only twenty-two percent of the between class variation in learner achievement. The three competencies in the model relate to organizing time, space and materials (Competency 4), demonstrating acceptable oral and written expression and knowledge of subject matter (Competency III), and communicating with learners (Competency V). These values contrast with the correlation of the total instrument score with the teacher effectiveness index, where r=.32.

Insert Tables 5 and 6 about here

The results of the generalizability analyses of the indicators are summarized in Table 7 and the analyses of the competencies and total instrument in Table 8. Values for rho squared for the indicators ranged from 0 (Indicator 14, demonstrates command of the subject, and Indicator 21, Uses instructional aids and materials) to .66 (Indicator 30, Manages disruptive behavior among learners). Values of phi(lamda) were substantially higher although several were as low as .4. Values of rho squared for the competencies were quite variable, ranging from .08 to .71 for competencies three and eight respectively. The basis for the low values can be seen in the variance components which show relatively low teacher effects in both cases

and a substantial teacher by observer type interaction in the case of Competency VI. Values for phi (lamda) were substantially higher, however, with the lowest being .53 for Competency III, demonstrates acceptable written and oral expression. Rho squared for the instrument was .71.

Insert Tables 7 and 8 about here

DISCUSSION

A successfully operating assessment program such as the initial certification program in Georgia limits a researcher's access to the teachers who might be affected by it. Teachers who are involved in an assessment probably should and can not be saddled with the added pressures of being research subjects. Consequently, this research was limited to experienced teachers. Of course, there was value in being able to include an entire set of teachers in this pilot study. The variability of teaching performance may have been greater in this sort of pilot study than it is with highly motivated beginning teachers who are selecting an "optimum" context. This increased variance may contribute to increased reliability and to the possibility of finding significant relationships with the criterion variable(s). However, these benefits to reliability may be offset by the fact that the observers may have been error prone since they were not experienced with the new instrument. The cumulative effects of these cannot be known, of course, but possible limitations should be acknowledged.

In Georgia The principal interest of work such as this lies with its potential implications for instrument credibility,

particularly validity and reliability. Since decisions are made about competencies in Georgia, these scores are of most pertinent. The fact that all but one competency was related to the effectiveness index can certainly be encouraging. Compared to the competencies, a smaller portion of indicators and descriptors were related to the effectiveness index and some correlations were negative. Nevertheless, even the proportion of significant ($p<.05$) correlations among the descriptors (about 25%) was substantially above chance.

The magnitude of these correlations can raise questions. One of these is "What expectations should there be for the magnitude of coefficients such as these?" Several years ago there was a suyggestion that correlations should be higher than .39 before they are considered noteworthy (Medley, 1977). However, Gage and has suggested that much more modest correlations are all that can be expected and that small portions of variance accounted for may be significance (1978). The possibility of attenuation by unreliability should be acknowledged, however. In the present study, for example, with rho squared values for competency scores ranging from .08 to .71, the "true" correlations with the effectiveness index might be considerably higher than the values that were reported (Range of r values is .21 to .41.)

A second, related question involves the stability of the results in a variety of contexts. Expecting the same descriptors to relate to teacher effectiveness in other studies is not realistic. Similarly, expecting the same indicators to relate to effectiveness may be unreasonable. After all, some of these will have occurred by chance. And, some of these specific behaviors

may not be equally applicable in each specific context. The expectation for every competency to relate the effectiveness may be more reasonable. Although these reliability estimates in this study are tentative; competencies as a set are more reliable than are the more specific, subordinant elements. Furthermore, because the competencies are aggregates of specific behaviors, there is a likelihood that the aggregate will be related to effectiveness in various contexts even though every constituent descriptor will not.

To some extent, these last points begin to bear on the theme of the paper: "How many performance criteria should there be?" There should be enough criteria to make the process credible. No teacher with a substantial weakness should be able to pass into a "high status" category. For example, no teacher who is weak in management should be able to be called meritorious or even minimally competent. Furthermore, each of these criteria must be credible--with believability, reliability and validity. These principles apply equally to assessment systems--TPAI, TADS, PMS, TTAS, Virginia's or Tennessee's.

The results are particularly pertinent to system of dichotomous, summative judgements such as the T.P.A.I. or TTAS. The results demonstrate that these types of jdugements can be used to make dependable decisions and the scores can be related to teacher effectiveness. The results also suggest that the most desirable criterion level may an intermediate aggregation unit--larger than descriptors but less than a total instrument score. The total number of descriptors scored "acceptable" was quite reliable and it accounted for approximately 10% of the

variation in effectiveness as it was defined. Without great sacrifices in phi (lambda) many competencies were as valid as the entire instrument in this regard. And, since the competencies do not measure the same instructional processes, having multiple criteria enhances the likelihood of better predictions of achievement.

The regression analyses demonstrate the increased prediction with scores on three individual competencies predicting twice as much variance in teacher effectiveness as did the total. Selected individual indicators can predict half or more of the teacher variation effectiveness. However, caution must be used in this regard. With so many possible independent variables, there is the possibility of finding significant combinations which are not meaningful. The combination presented in Figure 8 may be a fortuitous combination which illustrates this point or it may illustrate an additional "trap." Indicator 10 may reflect a dimension of planning. Indicator 12 may represent a "G" factor. Indicators 11 and 18 may represent two critical aspects of performance--informally assessing learner progress and providing feedback. All of this may be true. However, none of these assertions may be true. And, certainly no directionality can be claimed in the relationship despite the logic that says that reaching the minimal even in both monitoring and providing feedback enhances achievement.

A different, more comprehensive validation model would be required to support this assertion. Perhaps an attempt to preassess teachers and learners could be followed by an attempt to "engineer" improvements in the performances of selected teachers.

Thus, if changes would be documented and associated with differences in learner performance, there could be a strong case for causality claims about T.P.A.I. behaviors influencing achievement as well as for a stronger case for the validity of the scores.

The current study does have a number of elements which would be desirable in the first engineering study of that sort. The common post-test, common content, use of a complete set of teachers, common post-test, and use of class as unit of analysis are all important elements in reducing some of the random variations in the study. Should such a controlled study produce "desirable" results, then a more realistic, longer term research effort might be made.

CONCLUSION

The results of this study show that the aggregation of summative judgements which is used in the scoring of the T.P.A.I. can be a valid and reliable procedure. The evidence suggests that intermediate levels of scoring such as the T.P.A.I. competencies are more desirable than total instrument scores. While the total is more reliable, it is a less valid indicator of effectiveness. The results provide evidence to support the validity and reliability of competency scores as they are used in Georgia since validity and dependability coefficients were adequate. However, caution should be exercised in inferring causality of these teacher behaviors or learners outcomes based on these results, based on this or similar studies.

References

Gage, N.L. (1978). The yield of research on teaching. Phi Delta Kappan, 60(3), 229-35.

Capie, W. (1983). Methodological issues in combining teacher performance data for decision making. A paper presented at the annual meeting of the American Educational Research Association, Montreal, April.

Roadrangka, V., Padilla, M. & Yeany, R. (1983). The development and validation of the Group Test of Logical Thinking (GALT). Paper presented at the annual meeting of the National Association for Research in Science Teaching, Dallas, Texas, April.

Hambleton, R.K. (1984). Validating the test scores. A Guide to Criterion-References Test Construction, 199-230.

Capie, W. (1985). Coming of age: Systematic performance evaluation. Educational Measurement, 4(3), 13-16.

Cronin, L., & Capie, W. (1985). A study of the content validity of the Revised Teacher Performance Assessment Instruments. Paper presented at the annual meeting of the Eastern Educational Research Association, Virginia, February.

Cronin, L., & Padilla, M. (1986). The middle grades integrated process skill test. Paper presented at the annual meeting of the National Association for Research in Science Teaching, San Francisco.

Summary of TPAI Organization

Planning

I. PLANS INSTRUCTION TO ACHIEVE SELECTED OBJECTIVES

1. Specifies or selects learner objectives for lessons.
2. Specifies or selects learning activities.
3. Specifies or selects materials and/or media.
4. Plans activities and/or assignments which take into account learner differences.

II. OBTAINS INFORMATION ABOUT THE NEEDS AND PROGRESS OF LEARNERS

5. Specifies or selects procedures or materials for assessing learner performance on objectives.
6. Uses systematic procedures to assess all learners.

III. DEMONSTRATES ACCEPTABLE WRITTEN AND ORAL EXPRESSION AND KNOWLEDGE OF THE SUBJECT

7. Uses acceptable written expression.

Observation

IV. ORGANIZES TIME, SPACE, MATERIALS, AND EQUIPMENT FOR INSTRUCTION

8. Attends to routine tasks.
9. Uses instructional time efficiently.
10. Provides a physical environment that is conducive to learning.

11. Assesses learner progrsss during the lesson observed.

12. Uses acceptable written expression with learners.
13. Uses acceptable oral expression.
14. Demonstrates command of school subject being taught.

V. COMMUNICATES WITH LEARNERS

15. Gives explanations related to lesson content.
16. Clarifies explanations when learners misunderstand lesson content.
17. Uses learner responses or questions regarding lesson content.
18. Provides information to learners about their progress throughout the lesson.

Observation

VI. DEMONSTRATES APPROPRIATE INSTRUCTIONAL METHODS

19. Uses instructional methods acceptably.
20. Matches instruction to learners.
21. Uses instructional aids and materials during the lesson observed.
22. Implements activities in a logical sequence.

VII. MAINTAINS A POSITIVE LEARNING CLIMATE

23. Communicates personal enthusiasm.
24. Stimulates learner interest.
25. Demonstrates warmth and friendliness.
26. Helps learners develop positive self-concepts.

VIII. MAINTAINS APPROPRIATE CLASSROOM BEHAVIOR

27. Maintains learner involvement in instruction.
28. Redirects learners who are off-task.
29. Communicates clear expectations about behavior.
30. Manages disruptive behavior.

Figure 1.

BEST COPY AVAILABLE

COMPETENCY VIII: MAINTAINS APPROPRIATE CLASSROOM BEHAVIOR

Indicator 28: Redirects learners who are off-task.

Descriptors

a. Non-verbal techniques are used to redirect learners who are off-task.

b. Verbal techniques are used to redirect learners who are off-task.

c. Learners who are off-task are effectively redirected.

d. Techniques are used to maintain attention of learners who have been redirected.

OR

5. No persistent off-task behavior is observed.

Figure 2.

BEST COPY AVAILABLE

Table 1

Mean Performance Level for TPAI Descriptors

Indicator	Descriptors a	b	c	d
8	.85	.78	1.00	.58
9	.98	.73	.95	.98
10	.98	1.00	.23	.98
11	1.00	.65	.58	.63
12	.93	.98	.98	.98
13	1.00	1.00	1.00	1.00
14	.98	1.00	1.00	.90
15	.83	.98	.93	.73
16	.63	.98	.95	.90
17	.98	1.00	1.00	.40
18	.85	.88	.83	.75
19	.98	.95	.58	.73
20	.93	.83	.90	.88
21	1.00	.88	.98	.85
22	.20	.95	.97	.45
23	.88	.85	.78	.73
24	.33	.40	.63	.53
25	.98	1.00	.90	.93
26	.85	1.00	.65	.75
27	.98	.73	.78	.80
28	.73	.95	.70	.60
29	.95	.85	.33	.95
30	.83	.85	.85	.85

Table 2

Mean Scores for TPAI Indicators and Competencies

Competency	Indicator	Mean	S.D.	Mean	S.D.
IV				.75	.21
	8	.67	.30		
	9	.69	.34		
	10	.88	.22		
II	11	.79	.30	.79	.30
III				.88	.11
	12	.84	.21		
	13	.95	.17		
	14	.86	.18		
V				.72	.26
	15	.68	.34		
	16	.88	.21		
	17	.84	.23		
	18	.48	.30		
VI				.62	.27
	19	.52	.35		
	20	.62	.32		
	21	.77	.22		
	22	.53	.33		
VII				.62	.28
	23	.58	.34		
	24	.35	.35		
	25	.88	.21		
	26	.65	.33		
VIII				.66	.37
	27	.63	.33		
	28	.53	.38		
	29	.76	.33		
	30	.71	.36		

Table 3

Correlation between Achievement and Mean Legal Team Descriptor Scores

Indicator	Descriptor a	b	c	d
8	.24	.05	.05	.22
9	.38*	.39**	.42**	.13
10	.56**	.26*	.09	.38**
11	.16	-.06	-.25	-.05
12	.28*	.28*	.06	.12
13	.27*	.21	-.18	.00
14	-.26	-.03	-.03	.08
15	.02	.22	-.08	.15
16	-.06	.25	.28*	.05
17	.09	.22	.56**	.15
18	-.01	.11	.01	.11
19	-.18	.14	.24	.25
20	.23	-.12	-.18	.40**
21	.21	.28*	-.01	.36*
22	.12	.21	-.25	-.07
23	.06	.07	.25	.05
24	.07	-.04	-.10	.11
25	.30*	.22	.15	.02
26	.31*	.09	.30*	.15
27	.12	.22	.23	.27*
28	.27*	.19	.35*	.23
29	.29*	.33*	-.12	.14
30	.06	.14	.31*	.22

Table 4

Correlations Between TPAI Indicators and Competencies and Teacher Effectiveness Index

Competency	Indicator	Correlation
IV		.42*
	8	.13
	9	.34*
	10	.53**
II	11	.22
III		.31*
	12	.36*
	13	.15
	14	.03
V		.21
	15	.09
	16	.26*
	17	.19
	18	.12
VI		.29*
	19	.25
	20	.24
	21	.13
	22	.15
VII		.27*
	23	.18
	24	.11
	25	.15
	26	.31*
VIII		.33*
	27	.27*
	28	.28*
	29	.29*
	30	.25

Table 5

Results of Regression Analyses of Selected TPAI Indicators

Source	Type IV SS	p	ΔR^2	R^2
10	109.5	.0002	27.7	27.7
11	28.8	.0368	2.6	30.3
12	54.9	.0050	9.4	39.7
18	26.4	.0452	6.6	46.3
Model	184.8	.0002		46.3

Table 6

Regression Analyses of Selected Competencies Against Teacher Effectiveness Index

Source	Type IV Sum of Squares	p	ΔR^2	R^2
Competency 4	44.4	.030	16.9	16.9
Competency 3	12.8	.217	1.9	18.8
Competency 5	8.7	.326	3.3	21.1
Model		.035		21.1

Table 7

Reliability Coefficients and Variance Components for TPAI Indicators

Indicator	ρ^2	$\phi(\lambda)$	T	O	TO
8	.27	.67	.025	.001	.199
9	.56	.76	.065	.000	.152
10	.38	.52	.019	.000	.092
11	.59	.70	.053	.006	.110
12	.09	.36	.004	.010	.124
13	.49	.54	.014	.000	.043
14	.00	.29	.000	.000	.134
15	.54	.76	.063	.000	.160
16	.24	.43	.010	.000	.100
17	.21	.44	.011	.003	.122
18	.11	.77	.010	.000	.242
19	.50	.82	.063	.006	.186
20	.30	.71	.030	.000	.210
21	.00	.44	.000	.000	.120
22	.34	.78	.037	.006	.211
23	.44	.77	.050	.008	.192
24	.52	.89	.063	.000	.170
25	.25	.43	.011	.002	.098
26	.44	.73	.047	.001	.182
27	.45	.73	.049	.016	.175
28	.60	.84	.085	.000	.168
29	.64	.76	.069	.000	.116
30	.66	.80	.084	.000	.127

Table 8

Reliability Coefficients and Variance Components for TPAI Competencies and Total Score

Competency	ρ^2	$\phi(\lambda)$	variance components T	O	I	TO	TI	OI	TOI
3	.08	.53	.001	.001	.002	.000	.000	.001	.107
4	.49	.75	.024	.000	.011	.020	.012	.001	.128
5	.33	.71	.013	.000	.031	.041	.011	.001	.115
6	.10	.79	.004	.000	.009	.057	.023	.006	.145
7	.41	.81	.013	.000	.045	.023	.025	.003	.137
8	.71	.86	.060	.002	.008	.040	.012	.002	.107
Total	.71	.92	.021	.000	.021	.017	.015	.002	.139